

Занятие 5. Лекция

План

1. Возвращаем долги
2. Тестирование гипотез

Jackknife (=перочинный ножик)

- метод, позволяющий улучшить нахалеву оценку

На прошлом занятии обсуждали bias-variance-tradeoff

$$MSE(\hat{\theta}_n) = \mathbb{E}[(\hat{\theta}_n - \theta)^2] = \underbrace{Bias(\hat{\theta}_n)}_{=(\mathbb{E}\hat{\theta}_n - \theta)^2} + \underbrace{Var(\hat{\theta}_n)}_{=\mathbb{E}[(\hat{\theta}_n - \mathbb{E}\hat{\theta}_n)^2]}$$



- как уменьшить смещение?
 - jackknife
- как уменьшить разброс?
 - методы монте-карло

В теории хочется сделать так:

$$\hat{\theta}_n^0 = \hat{\theta}_n - \mathbb{E}\hat{\theta}_n d$$

Тогда получается $\mathbb{E}\hat{\theta}_n^0 = 0$, т.е. оценка несмещенная. Но как найти это матожидание?

Пример

$$\theta = \sigma^2: \quad \hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{X})^2 - \text{смещ. оц. для } \theta$$
$$\mathbb{E}\hat{\theta}_n = \frac{n-1}{n} \cdot \sigma^2$$

Хочется вычесть, но вычесть не можем:

$$\hat{\theta}_n^0 = \hat{\theta}_n - \frac{n-1}{n} \sigma^2 \quad \text{но мы не знаем } \sigma^2$$

Идея: вычесть два раза θ :

$$\hat{\theta}_n^0 = \hat{\theta}_n - \text{Bias}(\hat{\theta}_n)$$

$$\hat{\theta}_n^0 - \theta = \hat{\theta}_n - \theta - \text{Bias}(\hat{\theta}_n) \quad | \quad \mathbb{E}[\cdot]$$

$$\text{Bias}(\hat{\theta}_n^0) = \text{Bias}(\hat{\theta}_n) - \text{Bias}(\hat{\theta}_n) = 0$$

Значит $\hat{\theta}_n^0$ несмещенная. Но мы не знаем $\text{Bias}(\hat{\theta}_n)$. Поэтому берем оценку (она называется jackknife):

$$\widehat{\text{Bias}}(\hat{\theta}_n) = (n-1) \left(\frac{1}{n} \sum_i \hat{\theta}_{(-i)} - \hat{\theta}_n \right)$$

Тогда $\mathbb{E}[\widehat{\text{Bias}}(\hat{\theta}_n)] = \text{Bias}(\hat{\theta}_n) + O(\frac{1}{n^2})$

Пример

$$\theta = \sigma^2; \quad \hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$\mathbb{E} \hat{\theta}_n = \frac{n-1}{n} \sigma^2 \Rightarrow \text{Bias}(\hat{\theta}_n) = \mathbb{E} \hat{\theta}_n - \sigma^2 = -\frac{\sigma^2}{n}$$

Метод jackknife:

$$\begin{aligned} \mathbb{E}[\widehat{\text{Bias}}(\hat{\theta}_n)] &= (n-1) \left(\frac{1}{n} \sum_{i=1}^n (\hat{\theta}_{(-i)} - \hat{\theta}_n) \right) \\ &= (n-1) \left(\text{Bias}(\hat{\theta}_{(-1)}) - \text{Bias}(\hat{\theta}_n) \right) \\ &= (n-1) \left(-\frac{\sigma^2}{n-1} + \frac{\sigma^2}{n} \right) = \frac{-\sigma^2}{n} = \text{Bias}(\hat{\theta}_n), \text{ т.е.} \end{aligned}$$

Получаем несмещенную оценку!!!

Типовая ситуация

Есть члены порядка $1/n, 1/n^2$

$$\text{Bias}(\hat{\theta}_n) = \frac{a}{n} + \frac{b}{n^2} + \dots$$

Тогда

$$\mathbb{E}[\widehat{\text{Bias}}(\hat{\theta}_n)] \stackrel{a}{=} (n-1) \left(\frac{a}{n-1} + \frac{b}{(n-1)^2} - \frac{a}{n} - \frac{b}{n^2} \right) = \frac{a}{n} + O(\frac{1}{n^2})$$

В итоге получаем:

$$\begin{aligned} \hat{\theta}_n^0 &:= \hat{\theta}_n - \widehat{\text{Bias}}(\hat{\theta}_n) \\ \text{Bias}(\hat{\theta}_n^0) &= \mathbb{E} \hat{\theta}_n^0 - \theta = \text{Bias}(\hat{\theta}_n) - \mathbb{E}[\widehat{\text{Bias}}(\hat{\theta}_n)] = \\ &= -\frac{\sigma^2}{n} + \frac{\sigma^2}{n} = 0 \end{aligned}$$

В случае типовой ситуации получаем

$$\text{bias}(\hat{\theta}_n^0) = O(\frac{1}{n^2})$$

Статистические тесты

- основной инструмент в матстатистике, чтобы доказать факт по данным

Основные вещи

Есть выборка взятая из закона распределения:

$$X_1, \dots, X_n \sim \mathcal{P}_\theta$$

Проверяем гипотезу:

$$\mathcal{H}_0 : \theta = \theta_0$$

Нужно найти C - критическую область, т.ч.

$$\mathbb{P}_{\theta_0}\{(X_1, \dots, X_n) \in C\} = \alpha = 0.05$$

Если мы попали в это множество, то отвергаем гипотезу

Пример (Нераскрытые парашуты)

Тестируем парашют с номером i :

$$X_1, \dots, X_n = \begin{cases} 1, & \text{парашют раскрылся,} \\ 0, & \text{иначе,} \end{cases} \quad \begin{matrix} p = \theta \\ p = 1 - \theta \end{matrix}$$

Гипотеза:

$$\mathcal{H}_0 : \theta = 0.0001$$

Критическое множество:

$$C = \{X_1 + \dots + X_n \geq t_\alpha\}$$

(число нераскрытых парашутов больше альфа).

$$\begin{aligned} \mathbb{P}_{\theta_0}\{X_1 + \dots + X_n \geq t_\alpha\} &= \alpha \\ &= \sum_{k=\lceil t_\alpha \rceil}^n C_n^k \theta_0^k (1 - \theta_0)^{n-k} = \alpha \end{aligned}$$

Нужно решить это уравнения относительно α . Можно с помощью R. Но есть другой способ, спомощью ЦПТ:

$$\mathbb{P}_{\theta_0}\{X_1 + \dots + X_n \geq t_\alpha\} = \mathbb{P}_{\theta_0}\left\{ \underbrace{\frac{X_1 + \dots + X_n - n\theta_0}{\sqrt{n\theta_0(1 - \theta_0)}}}_{\sim N(0,1)} \geq \frac{t_\alpha - n\theta_0}{\sqrt{n\theta_0(1 - \theta_0)}} \right\} = \alpha$$

Значит,

$$\frac{t_\alpha - n\theta_0}{\sqrt{n\theta_0(1 - \theta_0)}} = z_{1-\alpha} \Rightarrow t_\alpha = n\theta_0 + z_{1-\alpha} \sqrt{n\theta_0(1 - \theta_0)}$$

Мы считаем количество раз, сколько раскрылся парашют, с числом t_α . Если больше, то отклоняем. Иначе **не отклоняется**, но не принимается.

Но тогда что мы принимаем?

Число α , которые было выше, называется ошибкой 1 рода. Еще есть ошибки второго рода

Ошибки второго рода

$$\mathbb{P}_{\theta_0}\{(X_0, \dots, X_n) \in C\} = \alpha = 0.05$$

Идеально:

$$\mathbb{P}_{\theta_1}\{(X_1, \dots, X_n) \in C\} = 1 - \beta$$

где β - малое число. Оно является вероятностью ошибки второго рода.

А может ли быть так, что и $\alpha = \beta = 0$?

Пример

X_1, \dots, X_n

H_0 : параш. раскрылся
 H_1 : параш. не раскрылся

$\mathcal{P}_\theta = \left\{ \begin{array}{cc} \mathbb{P}_0 & \mathbb{P}_1 \\ \uparrow & \uparrow \\ \text{норм} & \text{не норм} \end{array} \right\}$

Для этой задачи есть тест, для которого оба числа равны нулю

$$C = \{X_1 \in \mathbb{N} \cup \{0\}\}$$

Тогда:

$$\begin{aligned}
 \mathbb{P}_0\{X_1 = 0, 1, 2, \dots\} &= \sum_{k=0}^{\infty} \mathbb{P}\{X_1 = k\} = 1 \Rightarrow \alpha = 0 \\
 \mathbb{P}_1\{X_1 = 0, 1, 2, \dots\} &= 1 \Rightarrow \beta = 0.
 \end{aligned}$$

Но это математическая экзотика

Какой размер выборки брать для эксперимента?

Максимально прикладной вопрос (вспоминаем тинек). Можно решить с помощью чисел α, β .

Пример про парашюты (продолжение)

$$P_{\theta_0} \left\{ \frac{X_1 + \dots + X_n - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}} \geq \frac{t_\alpha - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}} \right\} = \alpha$$

$\sim N(0,1)$

$$P_{\theta_1} \left\{ \frac{X_1 + \dots + X_n - n\theta_1}{\sqrt{n\theta_1(1-\theta_1)}} \geq \frac{t_\alpha - n\theta_1}{\sqrt{n\theta_1(1-\theta_1)}} \right\} = 1 - \beta$$

$\sim N(0,1)$

Мы ранее вывели

$$\begin{cases} \frac{t_\alpha - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}} = z_{1-\alpha} \\ \frac{t_\alpha - n\theta_1}{\sqrt{n\theta_1(1-\theta_1)}} = z_\beta \end{cases}$$

Откуда (приравниваем t_α)

$$n\theta_0 + \sqrt{n\theta_0(1-\theta_0)} z_{1-\alpha} = n\theta_1 + \sqrt{n\theta_1(1-\theta_1)} z_\beta$$

$$n = \left(\frac{\sqrt{\theta_1(1-\theta_1)} z_\beta - \sqrt{\theta_0(1-\theta_0)} z_{1-\alpha}}{\theta_0 - \theta_1} \right)^2$$

Графическая иллюстрация ошибок первого и второго рода

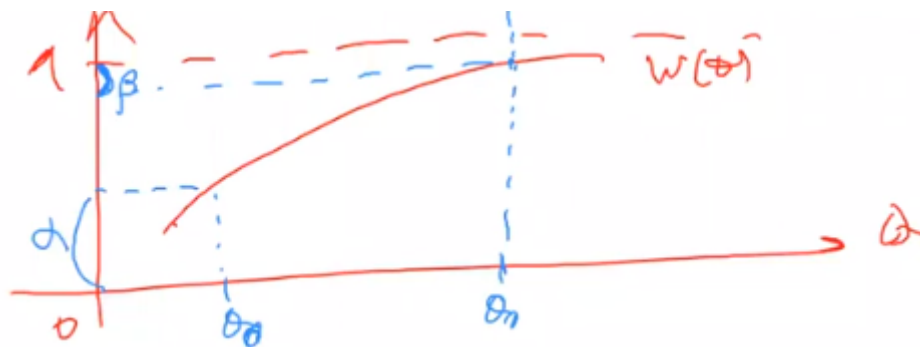
Функция мощности:

$$W(\theta) = P_\theta \{ (X_1, \dots, X_n) \in C \}$$

Типичный график:

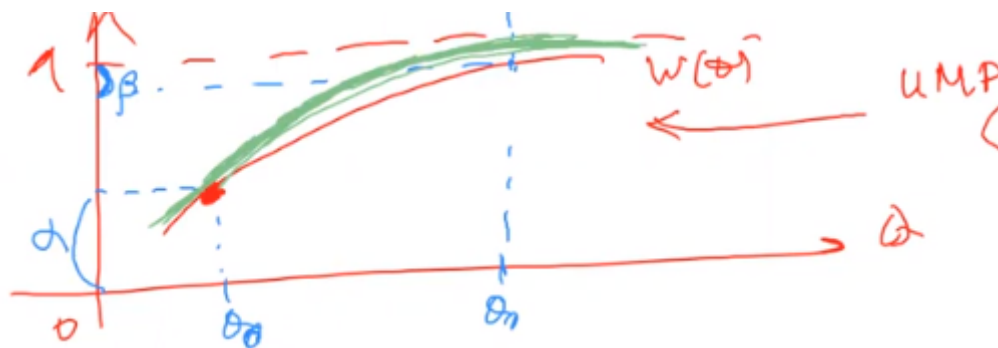


Вот так выглядят ошибки первого и второго рода:



- по параметру α понять как делать тест
- по параметру β можем понять для каких гипотез метод эффективный

Можно придумать тест, у которого α фиксирована, а кривая оптимальна? Да, это называется UMP tests (uniformly most powerful)



Такие тесты можно строить благодаря теореме Неймана Пирсона

UMP тесты

Теорема Неймана Пирсона

Есть две гипотезы $\theta = \theta_0$ и $\theta = \theta_1$. Тогда следующее критическое множество дает UMP тест:

$$C = \left\{ \frac{L_{\theta_1}(X_1, \dots, X_n)}{L_{\theta_0}(X_1, \dots, X_n)} > C_\alpha \right\}$$

где L - функция правдоподобия

Пример

$H_0: \mu = \mu_0$ vs $H_1: \mu = \mu_1$
 $X_i \sim N(\mu, \sigma^2)$
 $L_{\mu_0}(x_1, \dots, x_n) = \prod_{i=1}^n P(\mu, \sigma^2)(x_i) = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp \left\{ -\frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2 \right\}$
 $\frac{L_{\mu_1}(x_1, \dots, x_n)}{L_{\mu_0}(x_1, \dots, x_n)} = \exp \left\{ -\frac{1}{2\sigma^2} \left(\sum_{i=1}^n (x_i - \mu_1)^2 - \sum_{i=1}^n (x_i - \mu_0)^2 \right) \right\} \geq C_\alpha$
 $\mu_1 > \mu_0 \Rightarrow \sum x_i > \tilde{C}_\alpha$

Методика построения теста:

$$P_{\mu_0} \left\{ \sum_{i=1}^n X_i > \hat{c}_\alpha \right\} = \alpha$$

Тогда

$$\xi \sim N(0,1) \Rightarrow \left\{ \sum_{i=1}^n X_i > \hat{c}_\alpha \right\} = \left\{ \mu_0 n + n\sigma \xi > \hat{c}_\alpha \right\}$$

$$\left\{ \xi > \frac{\hat{c}_\alpha - \mu_0 n}{n\sigma} \right\}$$

(Федман-Пирсон)

И в итоге

$$\hat{c}_\alpha = \mu_0 n + n\sigma z_\alpha$$

$\sum X_i > \hat{c}_\alpha$ — откл. там
 $\sum X_i \leq \hat{c}_\alpha$ — не откл. там

И опять экспоненциальное распределение

Оно идеально подходит под теорию.

Мы свели неравенство из теоремы НП к неравенству с суммой X_i . В случае эксп. распр. так будет всегда

Экспоненциальное семейство:

$$p(x, \theta) = g(x) e^{x\theta - d(\theta)}$$

Посчитаем:

Статистический тест

Эксп. семейство:

$$p(x, \theta) = g(x) e^{x\theta - d(\theta)}$$

$$L_\theta(x_1, \dots, x_n) = g(x_1) \dots g(x_n) e^{\sum_{i=1}^n x_i \theta - n d(\theta)}$$

$$\frac{L_{\theta_1}(x_1, \dots, x_n)}{L_{\theta_0}(x_1, \dots, x_n)} = e^{\sum x_i (\theta_1 - \theta_0) - n(d(\theta_1) - d(\theta_0))} \stackrel{?}{>} c_\alpha$$

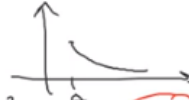
$$\theta_1 > \theta_0 \Rightarrow \boxed{\sum_{i=1}^n X_i > \hat{c}_\alpha}$$

Метод мощи UMP test: fix α ; $P_{\theta_0} \left\{ \sum X_i > \hat{c}_\alpha \right\} = \alpha$ — $\frac{\text{fact}}{\text{объем}} \hat{c}_\alpha$

Пример, когда лемма Немлана-Пирсона не работает

Пример, когда лемма Немлана-Пирсона не работает

$$f_{\theta}(x) = e^{-(x-\theta)} \mathbb{I}\{x \geq \theta\}$$



$$L_{\theta}(x_1, \dots, x_n) = \prod_{i=1}^n e^{-(x_i - \theta)} \mathbb{I}\{x_i \geq \theta\} = e^{-\sum x_i + n\theta} \mathbb{I}\{X_{(1)} \geq \theta\}$$

$$\frac{L_{\theta_1}(x_1, \dots, x_n)}{L_{\theta_0}(x_1, \dots, x_n)} = \begin{cases} e^{n(\theta_1 - \theta_0)}, & X_{(1)} \geq \theta_1 \\ 0, & \theta_0 \leq X_{(1)} < \theta_1 \end{cases}$$

т. Немлан-Пирсона: $\mathbb{P}_{\theta_0} \left\{ \frac{L_{\theta_1}(x_1, \dots, x_n)}{L_{\theta_0}(x_1, \dots, x_n)} > t_{\alpha} \right\} = \alpha$ — нужно решить относительно t_{α} .

Последнее уравнение решить невозможно:

нужно решить относительно t_{α}

$$\mathbb{P}_{\theta_0} \left\{ \frac{L_{\theta_1}(x_1, \dots, x_n)}{L_{\theta_0}(x_1, \dots, x_n)} > e^{-n(\theta_1 - \theta_0)} t_{\alpha} \right\} = \alpha$$

н.ч. = $\begin{cases} 0, & \tilde{t}_{\alpha} > 1 \\ \mathbb{P}_{\theta_0} \{X_{(1)} \geq \theta_1\}, & 0 < \tilde{t}_{\alpha} < 1 \end{cases}$ $\tilde{t}_{\alpha} = t_{\alpha} e^{-n(\theta_1 - \theta_0)}$

Вывод: $\frac{-n(\theta_1 - \theta_0)}{e} = d$ — тест можно построить.

$$(\mathbb{P}_{\theta_0} \{X_1 \geq \theta_1\})^n = (1 - \mathbb{P}\{X_1 < \theta_1\})^n = e^{-n(\theta_1 - \theta_0)}$$

Получили в конце странное равенство \Rightarrow тест построить не удастся

Позже обсудим, что делать

Рандомизированный статистический тест

- не является тестом в строгом смысле этого слова

Обычный тест:

$$\begin{aligned} \mathbb{P}_{\theta_0} \{(X_1, \dots, X_n) \in C\} &= \alpha \\ \Leftrightarrow \mathbb{E}[\mathbb{I}\{(X_1, \dots, X_n) \in C\}] &= \alpha \end{aligned}$$

Рандомизированный тест: вводим функцию d :

$$\begin{aligned} d(X_1, \dots, X_n) &\in [0, 1] \\ \mathbb{E}_{\theta_0}[d(X_1, \dots, X_n)] &= \alpha \end{aligned}$$

Интерпретация:

- $d(X_1, \dots, X_n) = 1$ - отклоняем
- $d(X_1, \dots, X_n) = 0$ - не отклоняем
- $d(X_1, \dots, X_n) \in (0, 1)$ - отклоняем с вероятностью $d(X_1, \dots, X_n)$

Аналог леммы Неймана-Пирсона:

$$d(X_1, \dots, X_n) = \begin{cases} 1, & \frac{L_{\theta_0}(X_1, \dots, X_n)}{L_{\theta_1}(X_1, \dots, X_n)} > t_\alpha \\ \gamma, & (\dots) = t_\alpha \\ 0, & (\dots) < t_\alpha \end{cases}$$

Тогда

$$1. \exists! \gamma, t_\alpha : \mathbb{E}_{\theta_0}[d(X_1, \dots, X_n)] = \alpha$$

2. Такой тест самый мощный среди всех тестов с вероятностью ошибки первого рода α , т.е.

$$\forall d^* : \mathbb{E}_{\theta_0}[d^*(X_1, \dots, X_n)] = \alpha \mid \mathbb{E}_\theta[d^*(X_1, \dots, X_n)] \leq \mathbb{E}_\theta[d(X_1, \dots, X_n)] \quad \forall \theta$$

Пример

Рандомизир. тест. $f_\theta(x) = e^{-(x-\theta)} \mathbb{1}\{x > \theta\}$

$$\frac{L_{\theta_1}(x_1, \dots, x_n)}{L_{\theta_0}(x_1, \dots, x_n)} = \frac{e^{n(\theta_1 - \theta_0)} \mathbb{1}\{x_{(1)} > \theta_1\}}{e^{n(\theta_1 - \theta_0)} \mathbb{1}\{x_{(1)} > \theta_1\}} > t_\alpha \quad \text{--- } C_\alpha$$

$$d(x_1, \dots, x_n) = \begin{cases} 1, & \mathbb{1}\{x_{(1)} > \theta_1\} > C_\alpha \\ \gamma, & \mathbb{1}\{x_{(1)} > \theta_1\} = C_\alpha \\ 0, & \mathbb{1}\{x_{(1)} > \theta_1\} < C_\alpha \end{cases} \quad \leftarrow \begin{matrix} C_\alpha = 0 \text{ или } C_\alpha = 1 \\ \Leftrightarrow \end{matrix}$$

$C_\alpha = 0$: $\mathbb{E}_{\theta_0}[d(x_1, \dots, x_n)] = \mathbb{E}[\mathbb{1}\{x_{(1)} > \theta_1\} + \gamma \mathbb{1}\{x_{(1)} \leq \theta_1\}]$
 $= P\{x_{(1)} > \theta_1\} + \gamma P\{x_{(1)} \leq \theta_1\} = e^{-n(\theta_1 - \theta_0)} + \gamma(1 - e^{-n(\theta_1 - \theta_0)}) = \alpha$
 $\gamma = \frac{\alpha - e^{-n(\theta_1 - \theta_0)}}{1 - e^{-n(\theta_1 - \theta_0)}} \in [0, 1] \Leftrightarrow e^{-n(\theta_1 - \theta_0)} < \alpha$

$C_\alpha = 1$: $\mathbb{E}_{\theta_0}[d(x_1, \dots, x_n)] = \mathbb{E}[\gamma \mathbb{1}\{x_{(1)} \leq \theta_1\}] =$
 $= \gamma \cdot P\{x_{(1)} \leq \theta_1\} = \gamma e^{-n(\theta_1 - \theta_0)} = \alpha$
 $\gamma = \alpha e^{n(\theta_1 - \theta_0)} \in [0, 1] \Leftrightarrow e^{-n(\theta_1 - \theta_0)} \geq \alpha$

Когда $(\dots) = \alpha$, то используем обычную лемму НП

LR-тесты

У нас были очень простые гипотезы: параметр равен одному числу или другому. Обычно тесты другие: параметр равен заданному числу или нет:

$$\mathcal{H}_0 : \theta \in \Theta_0 \subset \Theta$$

$$\mathcal{H}_1 : \theta \in \Theta_1 \subset \Theta$$

Очень хочется обобщить лемму НП:

$$\left\{ \frac{\max_{\theta \in \Theta_1} L_{\theta}(x_1, \dots, x_n)}{\max_{\theta \in \Theta_0} L_{\theta}(x_1, \dots, x_n)} > C \right\}$$

Класс таких тестов называется LR.

Самый важный результат в этой области:

Теорема Уилкса

Пусть $\Theta_0 \subset \Theta$. Тогда

$$2 \log \left(\frac{\max_{\theta \in \Theta} L_{\theta}(X_1, \dots, X_n)}{\max_{\theta \in \Theta_0} L_{\theta}(X_1, \dots, X_n)} \right) \xrightarrow{d} \chi^2_{\dim \Theta - \dim \Theta_0}$$

где $X_p^2 = \xi_1^2 + \xi_p^2$, $\xi_i \sim N(0, 1)$

У этой теоремы есть предположения, но опустим их (это условия регулярности). Они почти всегда выполнены.

Пример

упр $x_1, \dots, x_n \sim N(\mu, \sigma^2)$, σ^2 - изв. $H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$ $\xi_1, \dots, \xi_p \sim N(0, 1)$

$$L_{\mu}(x_1, \dots, x_n) = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n e^{-\sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2}}$$

$$\log \left(\max_{\mu} L_{\mu}(x_1, \dots, x_n) \right) = \max_{\mu} \left(-n \log(\sqrt{2\pi}\sigma) - \sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma^2} \right)$$

$$= -n \log(\sqrt{2\pi}\sigma) - \frac{\sum (x_i - \bar{x})^2}{2\sigma^2}$$

$$\log \left(\max_{\mu} L_{\mu}(x_1, \dots, x_n) \right) = -n \log(\sqrt{2\pi}\sigma) - \frac{\sum (x_i - \mu_0)^2}{2\sigma^2}$$

$$2 \log(\dots) = \frac{n}{\sigma^2} (\bar{x} - \mu_0)^2 \xrightarrow{d} \chi^2_1$$

В нашем примере тут вообще равенство, а не стремление

Теорема Пирсона (1900-ый год)

Теорема Пирсона $x_1, \dots, x_n \text{ i.i.d.}$ $\xi_1, \dots, \xi_p \sim N(0, 1)$

$D_1, D_2, D_3, \dots, D_K \rightarrow \bigcup_{i=1}^K D_i = \mathbb{R}$ $P\{X \in D_i\} = q_i, i=1, \dots, K$

$N_i = \{ \text{кол. во экз. } X \in D_i \}$

ув: $\sum_{i=1}^K \frac{(N_i - nq_i)^2}{nq_i} \xrightarrow{n \rightarrow \infty} \chi^2_{K-1}$

Прим.: $X \sim F \leftarrow \text{fixed } q_i, P\{X \in D_i\} = F(u_{i+1}) - F(u_i)$ χ^2_{K-1}

Трудности: 1) $\frac{(N_i - nq_i)^2}{nq_i}$ \xrightarrow{d} $\{u_i, u_{i+1}\}$ $\alpha: P\{Y > t_{\alpha}\} = \alpha$

Ивченко, Мезьжева (1984): $\{nq_i \geq 5 \forall i\}$ $\frac{\sum (N_i - nq_i)^2}{nq_i} > t_{\alpha}$ $n \geq 50$ мен. откл.

2) $F \not\sim F_0 \in \text{negl } \Phi$

Решение первой проблемы - эмпирические условия. Решение второй проблемы (через матан):

цель минимизация χ^2 : $q_i = F(u_{i+1}) - F(u_i) = q(\theta)$

$$\hat{\theta}_1 = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^K \frac{(N_i - u q_i(\theta))^2}{u q_i(\theta)} \Rightarrow \sum_{i=1}^K \frac{(N_i - u q_i(\theta))^2}{u q_i(\theta)} \xrightarrow{u \rightarrow \infty} \chi^2_{K-1-\dim \Theta}$$

Почему из теоремы Уилкса следует теорема Пирсона?

- просто полезная техника, характерная для этой области матанализа

Теорема Уилкса (Wilks theorem)

$$\Theta_0 \subset \Theta \quad 2 \log \left(\frac{\max_{\theta \in \Theta} L_\theta(x_1, \dots, x_n)}{\max_{\theta \in \Theta_0} L_\theta(x_1, \dots, x_n)} \right) \xrightarrow{d} \chi^2_{\dim(\Theta) - \dim(\Theta_0)}$$

$\chi^2_p: \sum_{i=1}^p z_i^2 + \sum_{p+1}^{\infty} z_i^2$
 $z_1, \dots, z_p \sim N(0,1)$

Теорема Пирсона $X_1, \dots, X_n \text{ i.i.d.}$

$D_1, D_2, D_3, \dots, D_K \rightarrow \bigcup_{i=1}^K D_i = \mathbb{R}$

ув: $\sum_{i=1}^K \frac{(N_i - u q_i)^2}{u q_i} \xrightarrow{u \rightarrow \infty} \chi^2_{K-1}$

$P\{X \in D_i\} = q_i, i=1, \dots, K$
 $N_i = \{ \text{кол. ло экз. } X \text{ выпавших в } D_i \}$

Док: $P\{N_1 = n_1, \dots, N_K = n_K\} = \frac{n!}{n_1! \dots n_K!} p_1^{n_1} \dots p_K^{n_K}$ - мультиплик. распр

$\Theta_0: p_i = q_i, \dim(\Theta_0) = 0$

$\Theta: p_1 + \dots + p_K = 1, \dim(\Theta) = K-1$
 $\log(\max_{\theta} L_\theta(x_1, \dots, x_n)) = \frac{n!}{n_1! \dots n_K!} \left(\frac{n_1}{n}\right)^{n_1} \dots \left(\frac{n_K}{n}\right)^{n_K}$

В конце использовали

$$\frac{\partial}{\partial p_i} (n_1 \log p_1 + \dots + n_K \log p_K - \lambda(p_1 + \dots + p_K - 1)) = \frac{n_i}{p_i} - \lambda \Rightarrow$$

$$p_i = \frac{n_i}{n}$$

$$\lambda = n$$

Ок, идем дальше. Подставляем в формулу из теоремы Уилкса:

$$2 \log \left(\frac{\left(\frac{n_1}{n}\right)^{n_1} \dots \left(\frac{n_K}{n}\right)^{n_K}}{q_1^{n_1} \dots q_K^{n_K}} \right) = 2 \sum_{i=1}^K n_i \log \left(\frac{n_i}{u q_i} \right)$$

Разложим в ряд Тейлора:

Ф. Тейлора: $x \log \left(\frac{x}{x_0} \right) = (x - x_0) \left(x \log \frac{x}{x_0} \right)' \Big|_{x=x_0} + \frac{1}{2} (x - x_0)^2 \left(x \log \frac{x}{x_0} \right)'' \Big|_{x=x_0} + \dots = (x - x_0) + \frac{(x - x_0)^2}{2x_0} + \dots$

$$\Rightarrow \sum_{i=1}^K \left(n_i - u q_i + \frac{(n_i - u q_i)^2}{2 u q_i} \right) = \sum_{i=1}^K \frac{(n_i - u q_i)^2}{u q_i} \quad \square$$

$\sum_{i=1}^K n_i = n, \sum_{i=1}^K q_i = 1$

С теоремой Пирсона будет работать на семинарах.

Критерий хи-кварат для таблиц сопряженности

Таблица сопряженности (contingency table):

регулярно нерегулярно	25-30	31-35	...
да			
нет			

Можно было бы использовать корреляцию Пирсона. Но это предполагает, что распределение величин нормальное

Критерий хи-кварат для таблиц сопряжен.
contingency table

регулярно нерегулярно	25-30	31-35	...
да	•	•	•
нет	•	•	•

$A_i = \{ \text{случ. вел } X: \text{осознает } \text{сб. вел } Y: i\text{-ой } Y \}$
 $B_j = \{ \text{случ. вел } X: \text{осознает } \text{сб. вел } Y: j\text{-ой } Y \}$
 $\Theta = \{ p_{ij}, i=1..n, j=1..m \}$ $\dim(\Theta) = nm - 1$
 $p_{ij} = P\{X \in A_i \cap B_j\}$ $\sum_{i,j} p_{ij} = 1$
 $A_i \perp B_j \forall i, j$
 $p_{ij} = P\{X \in A_i\} P\{X \in B_j\}$
 $\dim(\Theta)_0 = \underbrace{n-1}_{P\{X \in A_i\}} + \underbrace{m-1}_{P\{X \in B_j\}}$
 Т. Иаксе:
 $\dim(\Theta) - \dim(\Theta)_0 = nm - 1 - (n+m-2) = (n-1)(m-1)$

Непосредственно сам критерий:

$$\sum_{i=1}^n \sum_{j=1}^m \frac{(N_{ij} - E_{ij})^2}{E_{ij}} \xrightarrow{(n-1)(m-1)} \chi^2$$

$$E_{ij} = \frac{\sum_{k=1}^m N_{ik}}{n} \cdot \frac{\sum_{l=1}^n N_{lj}}{m}$$

expected N

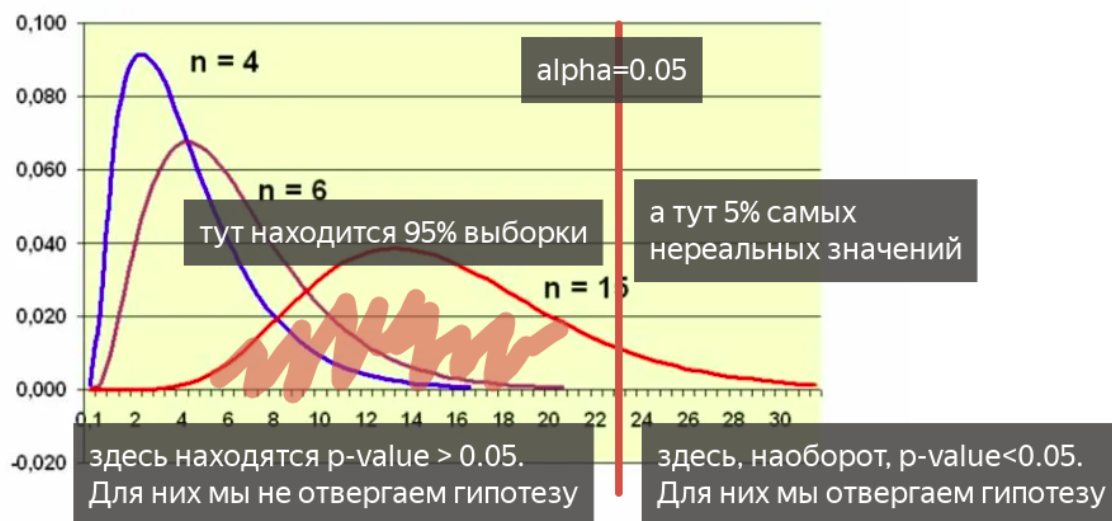
Пример

$$\begin{array}{cc|c} \textcircled{np} & E_{12} = \frac{12 \cdot 13}{22} & E_{12} = \frac{12 \cdot 9}{22} \\ & \textcircled{5} & \textcircled{7} & 12 \\ & E_{21} = \frac{10 \cdot 13}{22} & E_{21} = \frac{10 \cdot 9}{22} & 10 \\ & \textcircled{8} & \textcircled{2} & \\ 13 & 9 & | & 22 \end{array}$$
$$\chi^2_{(2-1)(2-1)} = \chi^2_1$$

Размышление про p-value

Если p-value меньше уровня значимости (например $\alpha = 0.05$, $p_{value} = 0.012$), то гипотеза отвергается. Иначе, если p-value больше уровня значимости, то гипотеза *не отвергается*, но и не принимается (считаем, что не хватает данных).

График хи-квадрат распределения



Вот снизу пример:

```
13 ~ }
14 p_hat=0.01*which.min(res)
15 1-pchisq(res[which.min(res)],df=4-1)
16 qchisq(0.95,df=3)
17

16:17 (Top Level) ↕

Console Terminal × Background Jobs ×

R 4.2.3 ~ /

+ p=j*0.01
+ E=N*choose(4,0:4)*(p^(0:4))*(1-p)^(4-0:4)
+ res[j]=sum((O-E)^2/E)
+ }
> which.min(res)
[1] 55
> p_hat=0.01*which.min(res)
> res[which.min(res)]
[1] 8.619665
> 1-pchisq(res[which.min(res)],df=4-1)
[1] 0.03479931
> qchisq(0.95,df=3)
[1] 7.814728
> |
```

У нас получилось значение 8.61966. Значение, соответствующее $p\text{-value}=0.05$ равно 7.814. Наше значение находится правее порога, поэтому мы отвергаем гипотезу.

Для $p\text{-value}$ похожая ситуация. **$p\text{-value}$ равно площади с правой стороны (площади правого хвоста)**. У нас она меньше 0.05 \Rightarrow наше значение лежит правее порога, в области нереалистичных значений \Rightarrow гипотеза отвергается.

$p\text{-value}$ удобнее, чем просто считать квантиль, т.к. мы сразу получаем уровень значимости.

К слову, еще бывают двухсторонние тесты. В них мы смотрим площадь не только правого хвоста, но и левого.

Выводы

- поговорили про тесты, рандомизированные тесты
- перешли к теореме Уилкса
 - если где-то видим сходимость к кси-квадрату, скорее всего корень лежит в этой теореме
 - поговорили про два применения
 - теорема Пирсона
 - таблица сопряженности